Title	Overview of L2 corpora and resources
Version	2.0
Author(s)	Jakob Lenardič, Therese Lindström Tiedemann, Darja Fišer
Date	20-04-2018
Status	For approval
Distribution	BoD, NCF, UI
ID	CE-2018-1202

Table of contents

1. Introduction

In the following report, we present an overview of second language (L2) corpora, primarily focusing on those that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). The report was conducted in several steps:

(i) manually searching the VLO and the national consortia with keywords like “learner corpus”, “academic writing”, and “longitudinal corpus”;

(ii) cross-referencing L2 corpora that are listed on the UCL[1] website with the VLO and the national repositories; and

(iii) input provided by CLARIN UI and NC coordinators as well as the participants of the Workshop on interoperability of L2 resources and tools, which took place in December 2017.

In total, more than 180 resources were identified but the majority seems to be unavailable through regular channels (i.e., download, concordancers), which is why in this report we focus on those that are available. In Section 2, we provide a comprehensive list of the L2 corpora that are part of the CLARIN infrastructure, and in Section 3 we describe their identification (i.e., listed in the VLO or not), their availability (download or through an online environment), and their metadata (size, annotation, target language, license). In Section 4, we provide a list of L2 corpora from languages spoken in CLARIN and that are available outside the CLARIN infrastructure. In Section 5, we provide a list of non-corpus resources related to L2 teaching.

2. L2 Learner Corpora in the CLARIN infrastructure

The second language (L2) learner corpora[2] consist mainly of written essays, often argumentative essays. But there are also some spoken corpora, some mixed and even one CMC corpus including material from a course taught over Adobe connect.

There degree of inclusion of L2 learner corpora into the CLARIN infrastructure is very good.

In this section we present the identified L2 learner corpora in the CLARIN infrastructure. We first list the monolingual corpora (1. written corpora; 2. spoken corpora; 3. Video and multimodal corpora.) and then the multilingual corpora in the same order.

2.1. Monolingual corpora

2.1.1. Written corpora

Table 1: Written monolingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description
CzeSL – Czech as a Second Language **Size:** 0.9 million words **Annotation:** tokenised, PoS-tagged, lemmatised, error labels **Licence:** CC-BY	Czech	This corpus contains essays written in 2013 by learners from 54 L1 backgrounds. The corpus is available for download from LINDAT. For a related publication, see Rosen (2016).

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[2] L2 is here used in its broad sense including foreign language, L2, L3, L4 etc. Any language learnt which is not an L1. Many corpora also tell you something of the informant's language learning background in the metadata.

British Academic Written English Corpus **Size:** 2761 texts **Licence:** CC-BY	English	This is primarily a L1 corpus although it also contains L2 texts. The corpus is available for download from the University of Oxford Text Archive.
CORYL (Corpus of Young Learner Language) **Size:** 191,568 tokens **Annotation:** tokenised, anonymised, error labels, linked to CEFR levels **Licence:** CC-BY	English	This corpus contains English texts written yb Norwegian primary school pupils (7th, 10th, and 11th grade). The corpus is available through the concordancer *Corpuscle* provided by CLARINO.
ETS Corpus of Non-Native Written English **Size:** 12,100 essays (1100 / language) **Licence:** restricted	English	The corpus contains texts written by learners from 11 L1 backgrounds as part of an international text of academic English proficiency. Prompts as well as proficiency level are part of the metadata. The corpus is available for download from the LDC catalogue.
The Hanken Corpus of Academic Writing **Size:** 500,000 words **Licence:** CC-BY	English	This corpus contains academic texts written by Finnish and Swedish native speakers. The corpus is still under development.
ICLE International Corpus of Learner English **Size:** 3 million words	English	This corpus contains texts written by learners of English from 14 L1 backgrounds. The corpus can be purchased on CD-ROM and a new version (ICLE v.3) is in development.
The Uppsala Student English corpus **Size:** 1.2 million tokens **Annotation:** tokenised **Licence:** CC-BY	English	This corpus contains essays written during the first three semesters of English studies at Uppsala University; most of the essays were written during the first semester. The corpus contains text files, each with a student ID and text ID including the course level, and information about the different prompts are available. The corpus is available for download from the University of Oxford Text Archive.
The Advanced Finnish Learners' Corpus **Size**: 288,000 tokens	Finnish	This corpus contains academic texts written by MA students and collected in 2009. The corpus consists of two subcorpora - The Exam Essays Subcorpus and the Course Papers

Annotation: tokenised, MSD-tagged, lemmatised **Licence**: CLARIN RES		Subcorpus, both of which are also available through Korp.
International Corpus of Learner Finnish (ICLFI) Corpus **Size:** 1 million words **Annotation**: MSD-tagged **Licence:** CLARIN RES	Finnish	This corpus contains fictional (e.g., letters, narratives) and non-fictional (e.g., essays) texts. The corpus provides information on a large number of variables concerning the linguistic background of the learner, the learning task, the learning context, etc. It is available through the concordancer Korp. For a related publication, see Jantunen (2011).
Testipiste Corpus **Size:** 840,000 tokens **Annotation**: tokenised **Licence:** CLARIN RES	Finnish	This corpus contains essays written by adult migrants from various L1 backgrounds. The corpus will be made available through the concordancer Korp.
Commented Learner Corpus Academic Writing **Size:** 853 texts **Licence:** CC BY-NC-SA 3.0	German	This corpus contains texts written by students at the University of Hamburg from various L1 backgrounds. The corpus is available for download through the repository of the University of Hamburg.
ASK – Norsk andrespråkskorpus **Size:** 618,000 tokens **Annotation**: tokenised, PoS-tagged, errors **Licence**: CLARIN RES	Norwegian	This corpus contains essays and tests written by students from 10 L1 backgrounds. It also contains L1 control essays. The corpus is available through a dedicated concordancer provided by CLARINO.
FinSveStud 79-80 **Size:** 175,000 tokens **Annotation**: tokenised, lemmatised **Licence:** CLARIN RES	Swedish	This corpus contains texts written by students with Finnish as their L1 background. The corpus is available through the concordancer Korp.
SW1203-essays **Size:** 52025 tokens **Annotation**: tokenised, PoS-tagged, MSD-tagged, lemgrams, compounds word forms **Licence:** CC-BY	Swedish	This corpus contains essays. The corpus is available for download through Språkbanken and through the concordancer Korp. Together with the Tisus corpus, SW1203-essays is a subcorpus of the pilot SweLL corpus

Tisus corpus **Size**: 60,000 tokens **Annotation:** tokenised, PoS-tagged, MSD-tagged, lemgrams, compounds word forms **Licence:** CC-BY	Swedish	Together with the SW1203-essays, the Tisus corpus is a subcorpus of the pilot SweLL corpus. The corpus is available for download from Språkbanken and through the concordancer Korp.

2.1.2. Spoken corpora

Table 2: Spoken monolingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description
The Anglish Corpus **Annotation:** interpausal units **Licence:** CLARIN RES	English	This corpus contains various speech tasks performed by French native speakers and the associated transcriptions. The corpus is available for download from Ortolang.
GLBCC (Giessen - Long Beach Chaplin Corpus) **Size:** 2472 words/transcript **Licence**: CC-BY	English	This corpus contains film retellings performed by English and German native speakers. The corpus is available for download from the University of Oxford Text archive.
ISLE Speech Corpus **Size:** approx. 18 hours **Annotation**: phone-level annotation, stress errors **Licence:** ELRA END USER	English	This corpus contains various speech tasks (reading simple sentences, using minimal pairs, giving answers to multiple choice questions) performed by German and Italian native speakers. The corpus is available for download from the ELRA catalogue.
A Learners' Corpus of Reading Texts **Licence:** CLARIN RES	English	This corpus contains unprepared readings by first-year students at an English department who speak French as a native language. The corpus is available for download from Ortolang.
French Learner Language Oral Corpora (FLLOC) **Size:**1375 transcripts **Annotation**: MSD-tagged **Licence**: CC-BY	French	This corpus contains various narrative and interactive speech tasks performed by English and Dutch native speakers. The corpus is available for download from the University of Oxford Text Archive.

Hamburg Modern Times Corpus **Size:** 24,000 words **Annotation**: prosody **Licence:** CLARIN RES	German	This corpus contains film retellings and the accompanying transcriptions. The corpus is available for download from the HZSK CLARIN-D repository.
LANGMAN **Size:** 11 subcorpora **Annotation:** error coding **Licence**: CC-BY	Hungarian	This corpus is a spoken corpus involving Chinese native speakers who learn Hungarian as a second language. The subcorpora are available for download from and browsing through the TalkBank.

2.1.3. Multimodal corpora

Table 3: Multimodal monolingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description
Arabic Learner Corpus **Size:** 0.3 million tokens **Annotation:** tokenised **Licence**: CLARIN RES	Arabic	This corpus contains essays written by students from 67 L1 backgrounds. It also contains recordings of speech tasks and associated transcriptions. The corpus is available for download from the LDC catalogue.
English as a Foreign Language Corpus **Size:** 24 hours **Licence**: Under Negotiation	English	The corpus contains videotaped lessons involving students at Finnish secondary schools.
The Long Second Corpus **Licence**: Under Negotiation	Finnish	This corpus contains written texts, audio recordings and videotaped lessons involving immigrants from the following L1 backgrounds: Estonian, Macedonian, Kurdish, Portuguese, Russian, and English. The corpus is still in preparation. It is set to be made available on the LAT platform.

2.2. Multimodal corpora

2.2.1. Written corpora

Table 4: Written multilingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description

CEFLING Project Corpus	Finnish and English	This corpus contains texts written by primary and secondary school students (years 7-9).
DIALUKI: Diagnosing reading and writing in a second or foreign language **Size:** 8,600 texts **Licence:** CLARIN RES	Finnish and English	This corpus contains texts both in Finnish (written by Russian native speakers) and English (written by Finnish native speakers). The corpus will be made available through Korp.
MERLIN Written Learner Corpus for Czech, German, Italian 1.1 **Size:** 2287 texts **Annotation:** a wide range of language characteristics that provide researchers with concrete examples of learner performance and progress across multiple proficiency levels. **Licence:** CC BY-SA 4.0	Czech, German, and Italian	This corpus contains learner texts produced in standardized language certifications covering CEFR levels A1-C1. The corpus is available for download from the Eurac Research CLARIN Centre Repository.
Topling - Paths in Second Language Acquisition **Size:**165,000 tokens **Annotation:** tokenised **Licence:** CLARIN End User Licence Agreement	Finnish, English and Swedish	This corpus contains written texts in English, Swedish and Finnish produced by students in the Finnish educational system and is an extension of the CEFLING corpus, which it also includes. The corpus is available through the concordancer Korp.

2.2.2. Spoken corpora

Table 5: Spoken multilingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description
AixOx **Size:** 40 minutes/task **Licence:** restricted	English and French	This corpus contains readings of written texts performed by French and English native speakers. The corpus is available for download from Ortolang.
	English and German	This corpus contains recordings of English and German spoken by non-native speakers from 31

LeaP: The Learning the Prosody of a Foreign Language **Size:** 31 hours **Annotation:** PoS-tagged, lemmatised, prosody		different native language backgrounds. The corpus is available for download from the Language Archive.
Repiso/Contrefactualité **Licence:** CLARIN RES	French, Italian, Spanish	This corpus contains recordings of counterfactual sentences. The corpus is available for download from ORTOLANG
Openprodat **Licence:** Publique Générale GNU	Dutch, English, French, German, Italian, Arabic, Spanish, Hungarian, Japanese, Thai, Norwegian, Chinese	This corpus contains paragraph readings by participants in both their L1 and in as many L2 as they felt they could manage. The corpus is available for download from Ortolang. For a related publication, see Hirst et al 2013
GeWiss **Size:** 1.4 million tokens **Annotation:** code switching	German (L2 and L1), English, Polish, Italian (L1)	This corpus contains transcripts and audio recordings of spoken academic discourse, primarily talks including discussions and oral exams. For the relevant publication, see Fandrych et al. (2014)

2.2.3. Multimodal corpora

Table 5: Multimodal multilingual L2-learner corpora in the CLARIN infrastructure

Corpus	Language	Description
TAITO: Written and Oral Data of the TAITO-project **Licence:** Under Negotiation	English, French, German, Italian, Swedish	This corpus contains texts written by undergraduate students at the beginning of their studies and videotaped discussions.
YKI National Certificates corpus **Licence:** CLARIN RES	Italian, Swedish, Spanish, English, Finnish, German, French, Russian	This corpus contains written and speech tasks.

3. Overview of the CLARIN L2-learner corpora

3.1. Identification

There are 36 L2 learner corpora that are integrated with the CLARIN infrastructure. All can be found in the VLO except for the following 2 corpora:

(1) SW1203-essays
(2) Tisus corpus

Both corpora are listed in the Språkbanken resource list, but not in the SWE-CLARIN repository.

3.2. Availability

The following 4 (11%) corpora are available for download and through a concordancer. In the parentheses, we list the consortium through which the corpus is available and, if applicable, the concordancer through which it can be queried.

(1) The Advanced Finnish Learners' Corpus (FIN-CLARIN; Korp)
(2) SW1203-essays (SWE-CLARIN; Korp)
(3) Tisus corpus (SWE-CLARIN; Korp)
(4) LANGMAN (TalkBank)

The following 6 (17%) corpora are available only through a concordancer:

(1) CORYL (Corpus of Young Learner Language) (CLARINO; Corpuscle)
(2) International Corpus of Learner Finnish (ICLFI) Corpus (FIN-CLARIN; Korp)
(3) ASK – Norsk andrespråkskorpus (CLARINO; Corpuscle)
(4) FinSveStud 79-80 (FIN-CLARIN; Korp)
(5) Topling - Paths in Second Language Acquisition (FIN-CLARIN; Korp)
(6) GeWiss (CLARIN-D; dedicated concordancer)

The following 17 (47%) corpora are available only for download:

(1) CzeSL – Czech as a Second Language (LINDAT)
(2) British Academic Written English Corpus (CLARIN-UK)
(3) ETS Corpus of Non-Native Written English (LDC Catalogue)
(4) The Uppsala Student English corpus (CLARIN-UK)
(5) Commented Learner Corpus Academic Writing (CLARIN-D)
(6) The Anglish Corpus (CLARIN-FR)
(7) GLBCC (Giessen - Long Beach Chaplin Corpus) (CLARIN-UK)
(8) ISLE Speech Corpus (ELRA)
(9) A Learners' Corpus of Reading Texts (CLARIN-FR)
(10)French Learner Language Oral Corpora (FLLOC) (CLARIN-UK)
(11)Hamburg Modern Times Corpus (CLARIN-D)
(12)Arabic Learner Corpus (LDC Catalogue)
(13)MERLIN Written Learner Corpus for Czech, German, Italian 1.1 (EURAC)
(14)AixOx (CLARIN-FR)
(15)LeaP: The Learning the Prosody of a Foreign Language (CLARIN-UK)
(16)Repiso/Contrefactualité (CLARIN-FR)
(17)Openprodat (CLARIN-FR)

The following 9 (25%) corpora are unavailable:

(1) The Hanken Corpus of Academic Writing (FIN-CLARIN)
(2) ICLE International Corpus of Learner English (FIN-CLARIN)
(3) Testipiste Corpus (FIN-CLARIN)
(4) The Long Second Corpus (FIN-CLARIN)
(5) DIALUKI: Diagnosing reading and writing in a second or foreign language (FIN-CLARIN)
(6) English as a Foreign Language Corpus (FIN-CLARIN)
(7) CEFLING Project Corpus (FIN-CLARIN)
(8) TAITO: Written and Oral Data of the TAITO-project
(9) YKI National Certificates corpus

Corpora (1)–(5) are under development. It is unclear why corpora (6)–(8) are unavailable. Corpus (9) seems to be available only for internal use.

3.3. Metadata

3.3.1. Languages covered

25 (69%) of the 36 corpora are monolingual. Among the monolingual corpora, there are 11 English, 4 Finnish, 3 Swedish, and 2 German corpora. There is only 1 corpus for each of the following languages: Arabic, Czech, French, Hungarian, and Norwegian.

3.3.2. Size

Information on size (i.e. tokens/number of documents or temporal length in case of video/audio corpora) is missing for the following corpora 8 (22%) out of 36 corpora:

(1) The Anglish Corpus
(2) A Learners' Corpus of Reading Texts
(3) The Long Second Corpus
(4) CEFLING Project Corpus
(5) Repiso/Contrefactualité
(6) Openprodat
(7) TAITO: Written and Oral Data of the TAITO-project
(8) YKI National Certificates corpus

The largest corpus in terms of token size is ICLE International Corpus of Learner English, which consists of 3 million tokens, while the smallest is Hamburg Modern Times Corpus, which consists of 24,000 tokens.

Otherwise, size is as follows:

- 12 small corpora (<1 million words/tokens)
- 4 medium-sized corpora (≥1 million words/tokens)

3.3.3. Annotation

Information on linguistic annotation is not available for the following 20 (56%) out of 36 corpora:

(1) British Academic Written English Corpus
(2) ETS Corpus of Non-Native Written English
(3) The Hanken Corpus of Academic Writing
(4) ICLE International Corpus of Learner English
(5) The Uppsala Student English corpus

(6) Testipiste Corpus
(7) Commented Learner Corpus Academic Writing
(8) GLBCC (Giessen - Long Beach Chaplin Corpus)
(9) A Learners' Corpus of Reading Texts
(10) Arabic Learner Corpus
(11) English as a Foreign Language Corpus
(12) The Long Second Corpus
(13) CEFLING Project Corpus
(14) DIALUKI: Diagnosing reading and writing in a second or foreign language
(15) Topling - Paths in Second Language Acquisition
(16) Repiso/Contrefactualité
(17) Openprodat
(18) TAITO: Written and Oral Data of the TAITO-project
(19) YKI National Certificates corpus
(20) AixOx

The annotation varies greatly from corpus to corpus. One of the text corpora is morphologically annotated (International Corpus of Learner Finnish (ICLFI) Corpus), while 2 text corpora display error annotation aside from tokenisation (ASK – Norsk andrespråkskorpus and CzeSL – Czech as a Second Language). Annotation of the audio corpora covers the markup of prosody errors (ISLE Speech Corpus), general prosody annotation (Hamburg Modern Times Corpus), general “interpausal units annotation” (ANGLISH) and 8-level annotation of pitch, tones, segments, syllables, words, phrasing, parts-of-speech and lemmata (LeaP: The Learning the Prosody of a Foreign Language).

10 (28%) corpora are PoS/MSD-tagged, while 5 (14%) corpora display error mark-up.

3.3.4. Licence

Information on licence is missing for the following 4 (11%) corpora:

(1) ICLE International Corpus of Learner English
(2) The Uppsala Student English corpus
(3) LeaP: The Learning the Prosody of a Foreign Language
(4) GeWiss

Otherwise, 12 (33%) corpora are available under CC-BY, 14 (39%) corpora are listed under restricted licences and the remainder under miscellaneous licences.

4. L2 Learner Corpora from CLARIN countries outside the CLARIN infrastructure

Apart from the L2 learner corpora included in the CLARIN infrastructure there are also many L2 learner corpora in the CLARIN countries which have not been included in the infrastructure for various reasons. Below we list those that we are aware of at the moment.

Table 6: Overview of L2 Learner corpora outside the VLO and/or national CLARIN repositories, but in countries within CLARIN

Corpus	Language	Description
ALEC: Advanced learner corpus of English **Size:** 1.3 million tokens	English (L1 mainly Swedish, but some other L1 and some L1 English)	This text corpus contains essays written by students of English (Uppsala University) in their third-fifth semester of studying English. Each file includes metadata about language background and year of study etc.
The Anglia Polytechnic University (APU) Learner Spanish Corpus **Size:** c. 120, 000 tokens	Spanish (L1 various)	This contains written compositions by learners of Spanish with different L1s. Annotation is unclear and it appears to be currently unavailable.
The ASU (andraspråkets strukturutveckling) corpus	Swedish	This corpus of Swedish L2 was compiled by Björn Hammarberg and uses the ITG system.
THE BAT MAT corpus: EFL academic writing	English (L1 Swedish, Finnish)	This corpus was created to study academic writing. It contains BA and MA dissertations of Finland-Swedish students at the University of Turku, Finland. There is background information about all writers however it unclear what the format of the corpus is and if it has been annotated.
CLC: The Cambridge Learner Corpus **Size:** 50 million tokens	English (L1 various)	This well-known corpus contains essays from learners of many different languages. The essays have been transcribed and error-annotated. Part of the CLC is available through SketchEngine. An earlier version of the corpus was described in a publication by D. Nicholls (2003).
EIC: The Estonian Interlanguage Corpus of Tallinn University	Estonian (L1 various)	This is a corpus of texts written by learners of Estonian as a second / foreign language. It can be searched online through a specific search interface. Annotation is unclear.

The Finnish as a Second Language (S2) Matriculation Examination Essay Collection 2001-2002 Corpus	Finnish (L1 Swedish?)	This is not yet included in the VLO. FIN-CLARIN states that negotiations about availability and licence are currently under way.
The Japanese learner corpus of Spanish **Size:** 83,000 tokens	Spanish (L1 Japanese)	It is unclear exactly what this corpus contains and whether it has been annotated. The corpus appears to be unavailable. Contact person: Yoshihito Kamakura
LANCAWE: The Lancaster Corpus of Academic Written English	English (L1 various)	This corpus consists of IELTS academic writing tests according to the Université catholique de Louvain list of L2 learner corpora. Weisser's list however claims that it consist of more varied writing from courses in academic L2 English. The corpus also contains a small control corpus of L1 data. Annotation is unclear however there are subcorpora according to task, writer, L1 etc. Contains longitudinal data.
Learner's corpus at the University of Tartu **Size:** 300,000 words	Estonian (L1 Various)	The information available about this corpus is only in Estonian and rather sparse. Contact person: kristiina.praakli@ut.ee
The LINCS Corpus	German (L1 English, German)	This corpus is under development according to the UCL list. Contact person: Elizabeth Thoday, Heriot-Watt University Edinburgh, UK.
Longi Corpus **Size:** 150,000 words	Swedish (L1 Finnish)	This consists of texts written by upper secondary students in Finland with Finnish as their L1. The corpus has been POS tagged. Licence: under negotiation (FIN-CLARIN)
The Longman Learners' Corpus **Size:** 10 million tokens	English (L1 various)	This corpus contains essays and exam scripts according to the UCL list. The UCL list it as a commercial, however availability is unclear.
The Lund CEFLE Corpus (Corpus Écrit de Français Langue Étrangère) **Size:** 100,000 words	French (L1: Swedish plus L1 French control group)	This text corpus contains essays written by Swedish learners of French (15-19 years old) as well as an L1 control group of the same age. The essays are available as text files without annotation. The corpus has been used within the Direkt profil project which provides some form of annotation but not quite clear which.

The Tartu Learner Corpus of Spanish as a L3+ **Size:** 885,000 words	Spanish (L1 Estonian)	This corpus contains texts written by Estonian-speaking learners of Spanish. Contact person: Mari Kruse, University of Tartu, Estonia
CLAG & GAP: Comasan Labhairt ann an Gàidhlig & Gaelic Adult Proficiency	Gaelic	This spoken corpus is based on a variety of oral tasks performed by the learners. It is unclear how it has been annotated and to whom it is available.
ProoF - Pronunciation of Finnish by Immigrants in Finland, version 1.0 **Size:** 20 hours	Finnish (L1 various)	This corpus contains read aloud and spontaneous speak by L2 speakers as well as an L1 control group. The corpus ha been transcribed and annotated on several levels in Praat (FIN-CLARIN).
SPLLOC: Spanish Learner Language Oral Corpus	Spanish (L1 English)	This corpus consists of recordings of learners of Spanish at different levelss as well as an L1 control group. The material contains narratives, interviews and picture description tasks. The corpus contains orthographic transcriptions and some error annotation. POS tagging and morphosyntactic tagging appears to have been done or be under way.

5. Textbook corpora in the CLARIN infrastructure

Apart from corpora of L2 learner texts there are also a few corpora of course book texts used in teaching L2 languages. These can be used to study the input that learners receive but also the expected receptive skills that they are expected to have at different proficiency levels.

So far we only know of two such corpora within the CLARIN countries, none of which are listed in the VLO, but the COCTAILL-corpus is available through Språkbanken's concordancer *Korp*.

Corpus	**Language**	**Description**
COCTAILL **Size:** 710,000 tokens	Swedish	This corpus is a collection of the reading texts in course books for Swedish as a second / foreign language at different levels. The corpus has been POS-tagged, syntactically annotated as well as annotated with CEFR-levels. More information in: Volodina et al 2014. Pilan et al 2016
Pedagogical Greek L2 textbooks corpus	Greek	This corpus is a compilation of texts from Greek L2 textbooks. The corpus consists of 276,000 tokens.

Size: 276,000 tokens		It is unclear how the corpus is annotated. The corpus is available for download from the clarin:el repository under the CC-BY licence. It is not listed in the VLO.

6. References

Hirst, Daniel, Brigitte Bigi, Hyongsil Cho, Hongwei Ding, Sophie Herment, Ting Wang. 2013. Building OMProDat: an open multilingual prosodic database. https://hal.archives-ouvertes.fr/hal-01510196.

Jantunen, Jarmo Harri. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. http://dx.doi.org/10.5128/LV21.04.

Orr, Rosemary, and Hugo Quené. 2017. D-LUCEA: Curation of the UCU Accent Project Data. https://doi.org/10.5334/bbi.15.

Rosen, Alexandr. 2016. Building and using corpora of non-native Czech. http://ceur-ws.org/Vol-1649/80.pdf.